No-Code Queries Can Accelerate AI and Data Analytics

By Dr. Jans Aasman, CEO

The low-code, no-code methodology is becoming highly sought-after throughout the modern IT ecosystem—and with good reason. Options that minimize manually writing code capitalize on the self-service, automation idiom that's imperative in a world in which working remotely and doing more with less keeps organizations in business.

Most codeless or low-code approaches avoid the need for writing language-specific code and replace it with a visual approach in which users simply manipulate on-screen objects via a drag-and-drop, point-and-click interface to automate code generation. The intuitive ease of this approach – which is responsible for new standards of efficiency and democratization of no-code development – has now extended to no-code query writing.

No-code querying provides two unassailable advantages to the enterprise. First, it considerably expedites what is otherwise a time-consuming ordeal, thereby accelerating data analytics and AI-driven applications and second, it can help organizations overcome the talent shortage of developers and knowledge engineers. Moreover, it does so by furnishing all the above benefits that make codeless and low-code options mandatory for success.

Read the full article at DZone.

Data-Centric Architecture Forum – DCAF 2021

Data and the subsequent knowledge derived from information are the most valuable strategic asset an organization possesses. Despite the abundance of sophisticated technology developments, most organizations don't have disciplines or a plan to enable data-centric principles.

DCAF 2021 will help provide clarity.

Our overarching theme for this conference is to **make it REAL**. Real in the sense that others are becoming data-centric, it is achievable, and you are not alone in your efforts.

Join us in understanding how data as an open, centralized resource outlives any application. Once globally integrated by sharing a common meaning, internal and external data can be readily integrated, unlike the traditional "application-centric" mindset predominantly used in systems development.

The compounding problem is these application systems each have their own completely idiosyncratic data models. The net result is that after a few decades, hundreds or thousands of applications implemented have given origin to a segregated family of disparate data silos. Integration debt rises and unsustainable architectural complexity abounds with every application bought, developed, or rented (SaaS).

Becoming data-centric will improve data characteristics of findability, accessibility, interoperability, and re-usability (FAIR principles), thereby allowing data to be exported into any needed format with virtually free integration.\

Dr. Jans Aasman to present – Franz's approach to Entity Event Data Modeling for Enterprise Knowledge Fabrics

---

Text Analytics Forum 2020 – KMWorld Connect

Join us November 17, 2020 – Text Analytics has the ability to add depth, meaning, and intelligence to any organization's most under-utilized resource – text. Through text analytics, enterprises can unlock a wealth of information that would not otherwise be available. Join us as we explore the power of text analytics to provide relevant, valuable, and actionable data for enterprises of all kinds.

Jans Aasman to present – Analyzing Spoken Conversations for Real-Time Decision Support in Mission-Critical Applications

November 17, 2020 at 2PM Eastern

Knowledge Graphs: A Single Source of Truth for the Enterprise

The notion of a “single source of truth” for the enterprise has been the proverbial moving goalpost for generations of CIOs, only to be waylaid by brittle technology and unending legacy systems. Truth-seeking visions rebuffed by technological trends have continuously confounded business units trying to achieve growth and market penetration. But technology innovation has finally led us to a point where CIOs can now deliver that truth.

Graphing the Truth

Knowledge graphs possess the power to deliver a single source of truth by linking together any assortment of data sources required, standardizing their diversity of data elements, and eliminating silos. They support the most advanced analytics options and decentralized transactions, which is why they’re now deployed as systems of records for some of the most significant, mission-critical use cases affecting our population.

Because they scale to include almost any number of applications – and link to other knowledge graphs as well – these repositories are the ideal solution for real-time information necessary to inform business users’ performances with concrete, data-supported facts. Most importantly, users can get an exhaustive array of touchpoints pertaining to any customer, product, or interaction with an organization from

the knowledge graph, making it a single source of truth.

Read the full article at Dataversity.

Using Microsoft Power BI with AllegroGraph

There are multiple methods to integrate AllegroGraph SPARQL results into Microsoft Power BI. In this document we describe two best practices to automate queries and refresh results if you have a production AllegroGraph database with new streaming data:

The first method uses Python scripts to feed Power BI. The second method issues SPARQL queries directly from Power BI using POST requests.

Method 1: Python Script:

Assuming you know Python and have it installed locally, this is definitely the easiest way to incorporate SPARQL results into Power BI. The basic idea of the method is as follows: First, the Python script enables a connection to your desired AllegroGraph repository. Then we utilize AllegroGraph's Python API within our script to run a SPARQL query and return it as a Pandas dataframe. When running this script within Power BI Desktop, the Python scripting service recognizes all unique dataframes created, and allows you to import the dataframe into Power BI as a table, which can then be used to create visualizations.

Requirements:

1. You must have the AllegroGraph Python API installed. If

you do not, installation instructions are here: https://franz.com/agraph/support/documentation/current/python/install.html

2. Python scripting must be enabled in Power BI Desktop. Instructions to do so are here: https://docs.microsoft.com/en-us/power-bi/connect-data/desktop-python-scripts

    a) As mentioned in the article, pandas and matplotlib must be installed. This can be done with 'pip install pandas' and 'pip install matplotlib' in your terminal.

The Process:

Once these requirements have been met, create a Python file with whatever script editor you usually use. The following code will create a connection to your desired repository. For this example, we will be using the Kennedy dataset that is available with the AllegroGraph distribution (See the 'Tutorial' directory). Load the Kennedy.ntriples file into your running AllegroGraph. (Replace the '****' in the code with your corresponding username and password.)

#the necessary imports

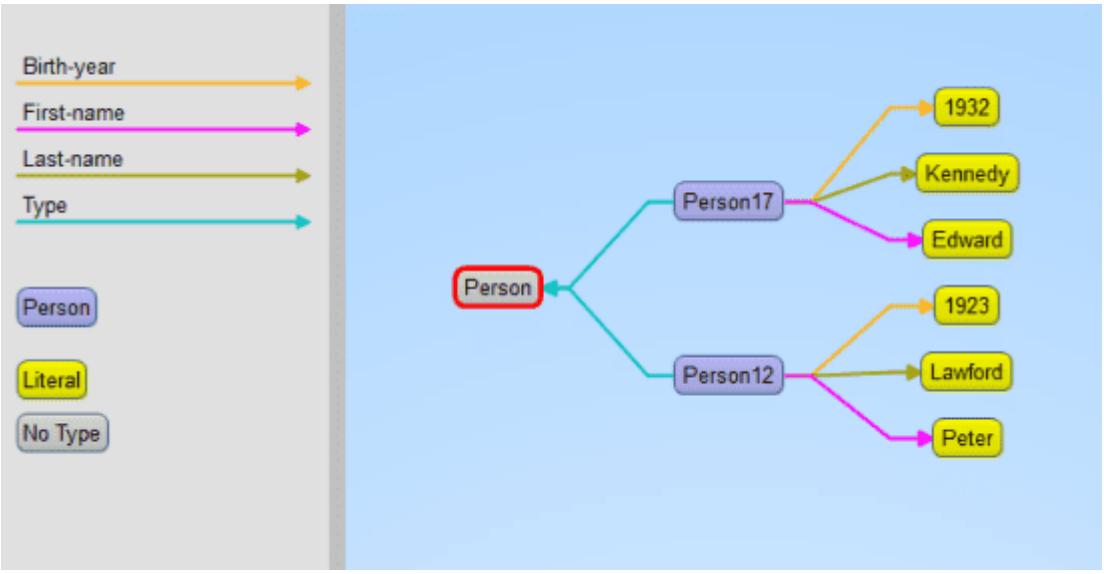
```
import os

from franz.openrdf.connect import ag_connect

from franz.openrdf.query.query import QueryLanguage

import pandas as pd
```

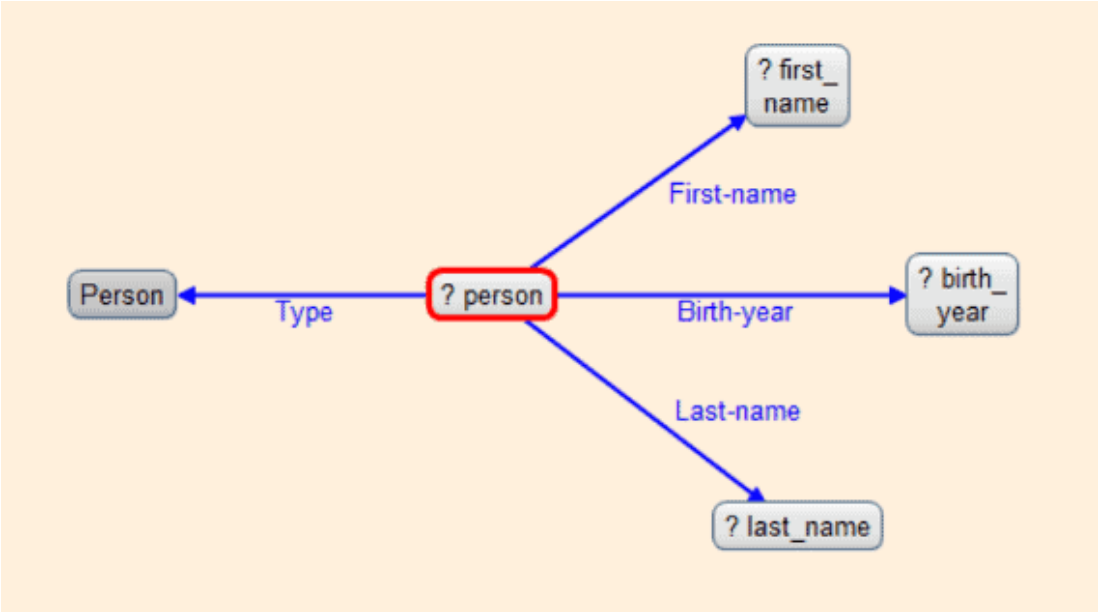
#connect to your agraph repository

```
def setup_env_var(var_name, value, description):

os.environ[var_name] = value
```

```
print("{}: {}".format(description, value))

setup_env_var('AGRAPH_HOST', 'localhost', 'Hostname')

setup_env_var('AGRAPH_PORT', '10035', 'Port')

setup_env_var('AGRAPH_USER', '****', 'Username')

setup_env_var('AGRAPH_PASSWORD', '****', 'Password')

conn = ag_connect('kennedy', create=False, clear=False)
```

2. We then want to create a query. For this example, we will first show what our data looks like, what the visual query of the information is, and what the written query looks like. With the following query we want every person's first and last names, as well as their birth years. Here is a small portion of the data visualized in Gruff, and then the visualization of the query:





3. Then add the written query to the python script as a variable string (we added an additional line to the query to sort on birth year). Next use the API functionality to simply execute the query and turn the results into a pandas dataframe.

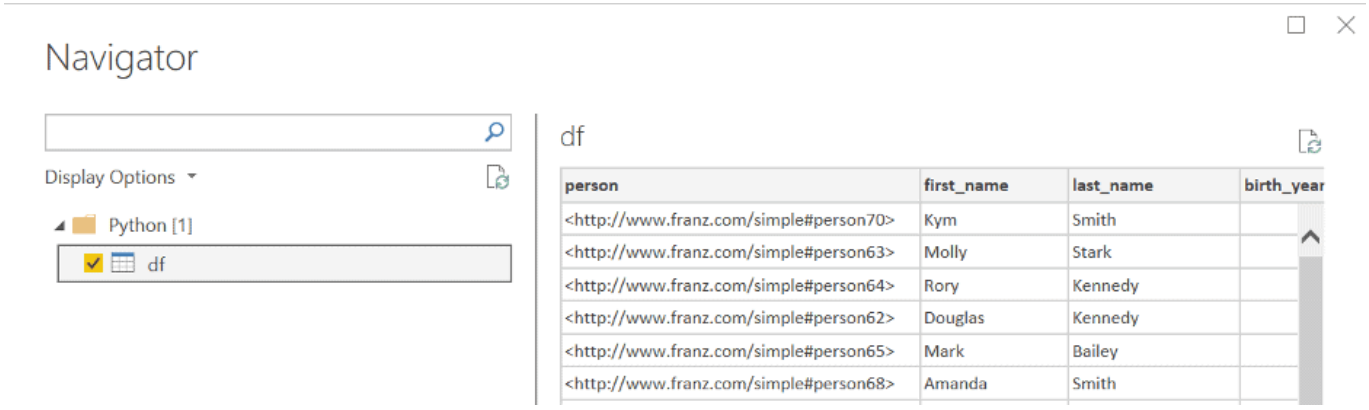
```
query = """select ?person ?first_name ?last_name ?birth_year
where
{ ?person <http://www.franz.com/simple#first-name> ?first_name
;
          <http://www.franz.com/simple#birth-year> ?birth_year
;
          rdf:type <http://www.franz.com/simple#person> ;
          <http://www.franz.com/simple#last-name> ?last_name .
}
order by desc(?birth_year)"""

with conn.executeTupleQuery(query) as result:
   df = result.toPandas()
```

When looking at the result, we see that we have a DataFrame!

```
[5] df.head()
```

	person	first_name	last_name	birth_year
0	<http://www.franz.com/simple#person70>	Kym	Smith	1972
1	<http://www.franz.com/simple#person63>	Molly	Stark	1968
2	<http://www.franz.com/simple#person64>	Rory	Kennedy	1968
3	<http://www.franz.com/simple#person62>	Douglas	Kennedy	1967
4	<http://www.franz.com/simple#person65>	Mark	Bailey	1967

4. Now we will use this script in Power BI. When in Power BI Desktop, go to 'Get Data' and look for the python script option. Then simply copy and paste your entire script into the text box, and run the script. In this case, our output looks like this:



5. Next simply 'Load' the data, and then you can use the Power BI Desktop interface to create whatever visualizations you want! If you do have a lot of additional operations to perform on your dataframe, we recommend doing these in your python script.

Method 2: POST Request:

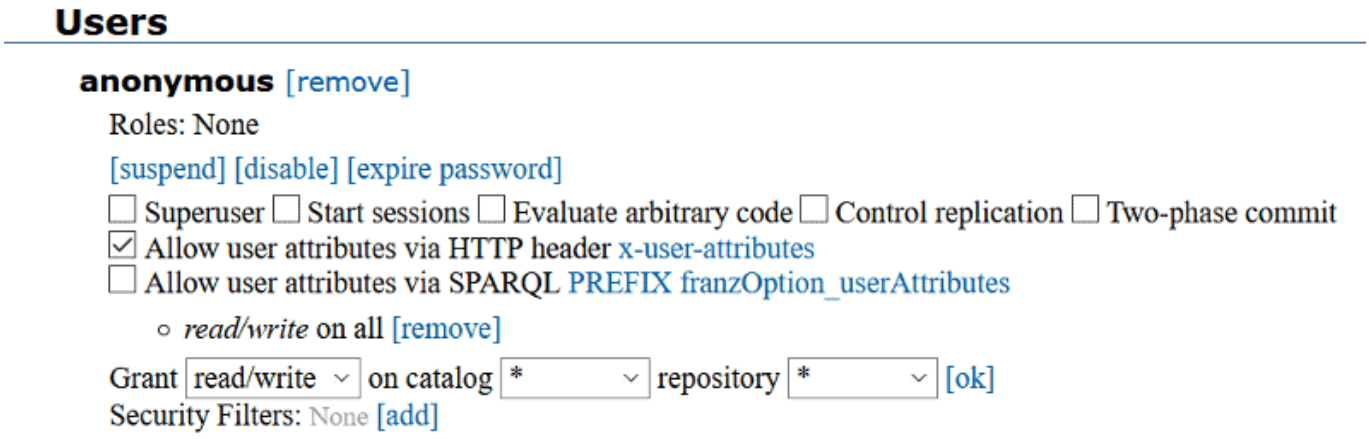
For the SPARQL query via POST requests to work you need to url-encode the query. Every modern programming language will support that, but in our example we will be using Python

again. This method is better for when you do not have python locally installed or prefer a different programming language.

It is possible to send a GET request from Power BI, but once the results from the query reach a certain size, a POST request is required, which is confusing to do within the Power BI Desktop interface. The following steps will show you how to do SPARQL Queries using POST requests. It looks a bit odd but it works well.

The Process:

1. In your AG WebView create an 'anonymous' user. (Go to admin -> Users -> [add a user] -> and add 'anonymous' as username without adding a password). You can use these settings:



2. Go to your desired repository in WebView and Click on 'Queries' -> 'New'

3. Write a simple SPARQL query, and run it to make sure you get the correct response back.

4. In python create the following script: (Assuming your AllegroGraph is on your localhost port 10035 and your repo is called 'kennedy')

```
import urllib
```

```
def CreatePOSTquery(query):
                                                start          =
"http://anonymous:@localhost:10035/repositories/kennedy?queryL
n=SPARQL&limit=1000&infer=false&returnQueryMetadata=false&chec
kVariables=false&query="
    response = start + urllib.parse.quote(query)
    return response
```

This function url-encodes the query and attaches it to the POST request. Replace the ‘localhost:10035’ and ‘kennedy’ strings in the start variable with your corresponding data. Then, using the same query as our previous example, we create our url-encoded POST query:

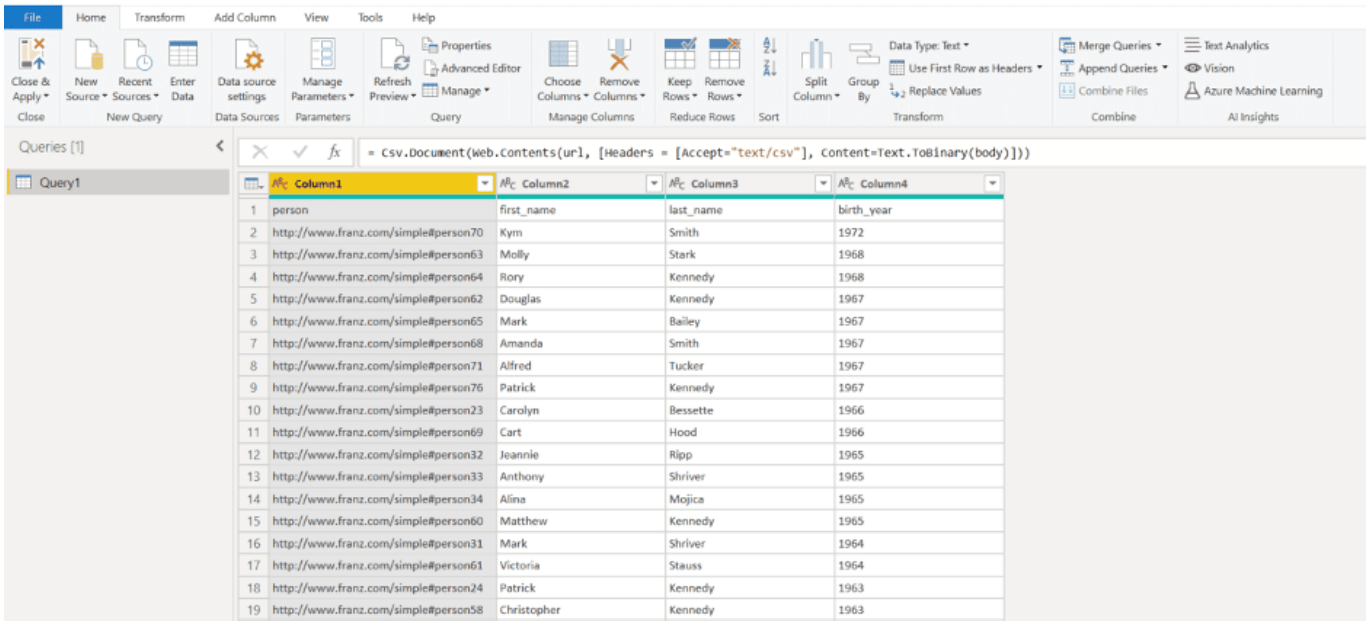
```
query = """select ?person ?first_name ?last_name ?birth_year
where
{ ?person <http://www.franz.com/simple#first-name> ?first_name
;
          <http://www.franz.com/simple#birth-year> ?birth_year
;
          rdf:type <http://www.franz.com/simple#person> ;
          <http://www.franz.com/simple#last-name> ?last_name .
}
order by desc(?birth_year)"""

result = CreatePOSTquery(query)
print(result)
```

This gives us the following result:

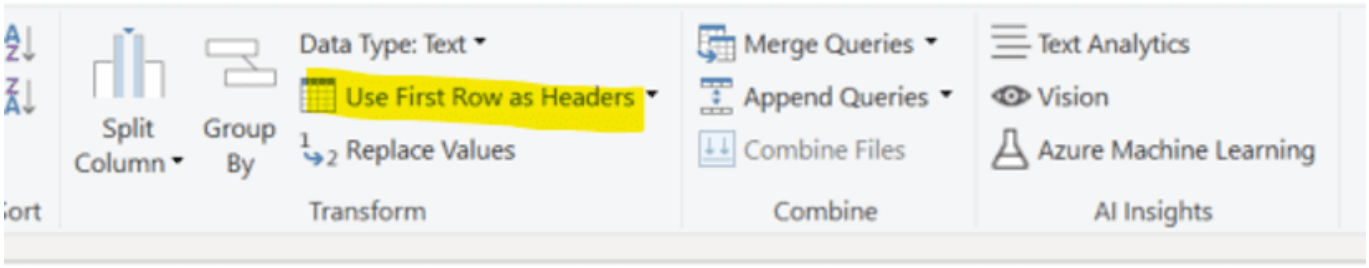
```
[16] result

'http://anonymous:@localhost:10035/repositories/kennedy?queryLn=SPARQL&limit=1000&infer=false&returnQueryMetadata=false&checkVariab
les=false&query=select%20%3Fperson%20%3Ffirst_name%20%3Flast_name%20%3Fbirth_year%20where%0A%7B%20%3Fperson%20%3Chttp%3A//www.fran
z.com/simple%23first-name%3E%20%3Ffirst_name%20%3B%0A%20%20%20%20%20%20%20%20%20%20%3Chttp%3A//www.franz.com/simple%23birth-year%3
E%20%3Fbirth_year%20%3B%0A%20%20%20%20%20%20%20%20%20%20rdf%3Atype%20%3Chttp%3A//www.franz.com/simple%23person%3E%20%3B%0A%20%20%2
0%20%20%20%20%20%20%20%3Chttp%3A//www.franz.com/simple%23last-name%3E%20%3Flast_name%20.%20%7D%0Aorder%20by%20desc%28%3Fbirth_year%
29'
```

5. Within Power BI Desktop we go to ‘Get data’ and create a ‘Blank query’ and go into the ‘Advanced Editor’ window. Using the following format we will get our desired results (please note that due to the length of the url-encoded request, it did not all fit in the image. Copy and pasting into the url field works fine. The ‘url’ variable needs to be in quotes and have a comma at the end):

We see the following results:



6. One last step is to turn the top row into the column names, which can be achieved by pressing the ‘Use first row as headers’:



The best part about both of these methods is that once the query has been created, Power BI can refresh the visuals using the same queries if your data changed. This can be achieved by scheduling refreshes within the Power BI Desktop interface (https://docs.microsoft.com/en-us/power-bi/connect-data/refresh-data#configure-scheduled-refresh)

Please send any questions or issues to: support@franz.com

Knowledge graphs enhance customer experience through speed and accuracy

KMWorld's recent article covers AllegroGraph and Franz's customer N3 Solutions.

The Full Article – KMWorld

KMWorld

Knowledge graphs are a way to model enterprise knowledge and

represent complex interrelationships in data. Information stored in a graph database can enable rapid retrieval of well-targeted results and provide insights into customers' interests and needs. Gartner predicts a 100% per-year growth in applications for graph analytics and databases for the next several years. Although knowledge graphs have been deployed by major companies such as Google, Amazon, and LinkedIn due to their ability to incorporate relationships in their analyses as well as their speed, only in the last 5 years has their use become more widespread.

N3 is an outsourced sales company for major organizations that sell complex B2B software, hardware, and tech solutions. It supports businesses in 92 countries, provides services in 25 languages, and holds thousands of hours of conversations every month with customers and prospects. "In today's world of complex products, it takes a well-educated team to tell the story about how this technology can help a company become more competitive," said Shannon Copeland, COO of N3. "The sales team needs to be able to instantly access the information they need to do their job."

Faster insights

The company has been operating for 16 years, and in the last few years began an initiative to manage its knowledge in a more intentional way. "We generate a great deal of data," noted Copeland, "and we wanted to make more effective use of it to understand our customers. And because of the speed at which business is transacted now, we needed to get insights right away, not a month later in a report."

N3 built a data model to reflect the essential data elements and the associations among them and decided that a knowledge graph was the best way to represent the information. After looking into partner options, N3 chose to work with Franz, Inc., which provides a semantic graph database called AllegroGraph. "We decided to work with Franz because of its

extensive experience and the fact that it had worked with a variety of industries," Copeland said.

The system built by N3 allows sales teams to organize signals from the market in a way that allows them to better explain the products to prospective buyers. "We build relationships with tech buyers on behalf of our clients," continued Copeland. "Our employees are typically college graduates who would like to begin their careers in sales and marketing in tech solutions. They take ownership of their territory and we help them be as sophisticated as a future CMO would be." The resources supplied by the knowledge graph provide the support the sales team needs to tailor information to each prospective customer.

The specific expertise required by the team varies depending on the products being sold, the geographic region, and other factors, and the knowledge graph supports these needs. For example, if a team in southern Portugal needs to know the preferences of that market, the associations built into the graph database can provide the information that is essential for them. "The information we can access helps customers understand the answers to their questions very quickly," Copeland commented. "We believe the experience that the customers have helps them scope out what they need and what the road map might be."

The strength of graph databases

A graph databases is a type of NoSQL database that stores data according to associations among data elements rather than in the rows and columns of a relational database. Because graph databases use a dynamic schema rather than a fixed, table-based one, adding new data types and categories is much easier. And because they are semantics-based, graph databases have strengths in inferring intent, producing answers to questions, and making recommendations. They can also make inferences about possible associations from existing

associations.

A graph database also provides much more context than a relational database and therefore can return more relevant results when a user is searching; they also integrate data from multiple sources. “At one telecom company we worked with, customer service reps might have [had] to open 15 databases to find out what went wrong and what the solution was,” said Jans Aasman, CEO of Franz. “We took their core customer data, billing information, every CRM call, and every action and put them into AllegroGraph, and the customer service reps were finally able to respond in a meaningful way, whether that was to make an offer to the customer or provide appropriate technical support.” The capability of graph databases to overcome silos and provide an integrated view of the customer is one of its strengths.

In order to create the graph database on which the knowledge graph is built, the relationships among entities need to be mapped. In the case of a hospital patient, the patient is the core entity, and the events are medical encounters or lab results, which may come out of different databases or a data warehouse. “The mapping is a major project, but it only needs to be done once,” Aasman pointed out. “After that, the relationships do not need to be regenerated during the search because they are indexed in AllegroGraph, which makes retrieval very rapid.”

AllegroGraph Named to 100 Companies That Matter Most in

Data

Franz Inc. Acknowledged as a Leader for Knowledge Graph Solutions

Lafayette, Calif., June 23, 2020 – Franz Inc., an early innovator in Artificial Intelligence (AI) and leading supplier of Semantic Graph Database technology for Knowledge Graph Solutions, today announced that it has been named to The 100 Companies That Matter in Data by Database Trends and Applications. The annual list reflects the urgency felt among many organizations to provide a timely flow of targeted information. Among the more prominent initiatives is the use of AI and cognitive computing, as well as related capabilities such as machine learning, natural language processing, and text analytics. This list recognizes companies based on their presence, execution, vision and innovation in delivering products and services to the marketplace.

“We’re excited to announce our eighth annual list, as the industry continues to grow and evolve,” remarked Thomas Hogan, Group Publisher at Database Trends and Applications. “Now, more than ever, businesses are looking for ways transform how they operate and deliver value to customers with greater agility, efficiency and innovation. This list seeks to highlight those companies that have been successful in establishing themselves as unique resources for data professionals and stakeholders.”

“We are honored to receive this acknowledgement for our efforts in delivering Enterprise Knowledge Graph Solutions,” said Dr. Jans Aasman, CEO, Franz Inc. “In the past year, we have seen demand for Enterprise Knowledge Graphs take off across industries along with recognition from top technology analyst firms that Knowledge Graphs provide the critical foundation for artificial intelligence applications and predictive analytics.

Our recent launch of AllegroGraph 7 with FedShard, a breakthrough that allows infinite data integration to unify all data and siloed knowledge into an Entity-Event Knowledge Graph solution will catalyze Knowledge Graph deployments across the Enterprise."

Gartner recently released a report "How to Build Knowledge Graphs That Enable AI-Driven Enterprise Applications" and have previously stated, "The application of graph processing and graph databases will grow at 100 percent annually through 2022 to continuously accelerate data preparation and enable more complex and adaptive data science." To that end, Gartner named graph analytics as a "Top 10 Data and Analytics Trend" to solve critical business priorities. (*Source: Gartner, Top 10 Data and Analytics Trends, November 5, 2019)*.

"Graph databases and knowledge graphs are now viewed as a must-have by enterprises serious about leveraging AI and predictive analytics within their organization," said Dr. Aasman "We are working with organizations across a broad range of industries to deploy large-scale, high-performance Entity-Event Knowledge Graphs that serve as the foundation for AI-driven applications for personalized medicine, predictive call centers, digital twins for IoT, predictive supply chain management and domain-specific Q&A applications – just to name a few."

Forrester Shortlists AllegroGraph

AllegroGraph was shortlisted in the February 3, 2020 Forrester Now Tech: Graph Data Platforms, Q1 2020 report, which recommends that organizations "Use graph data platforms to accelerate connected-data initiatives." Forrester states, "You can use graph data platforms to become significantly more productive, deliver accurate customer recommendations, and quickly make connections to related data."

Bloor Research covers AllegroGraph with FedShard

Bloor Research Analyst, Daniel Howard noted “With the 7.0 release of AllegroGraph, arguably the most compelling new capability is its ability to create what Franz refers to as “Entity-Event Knowledge Graphs” (or EEKGs) via its patented FedShard technology.” Mr. Howard goes on to state “Franz clearly considers this a major release for AllegroGraph. Certainly, the introduction of an explicit entity-event graph is not something I’ve seen before. The newly introduced text to speech capabilities also seem highly promising.”

AllegroGraph Named to KMWorld’s 100 Companies That Matter in Knowledge Management

AllegroGraph was also recently named to KMWorld’s 100 Companies That Matter in Knowledge Management. The KMWorld 100 showcases organizations that are advancing their products and capabilities to meet changing requirements in Knowledge Management.

Franz Knowledge Graph Technology and Services

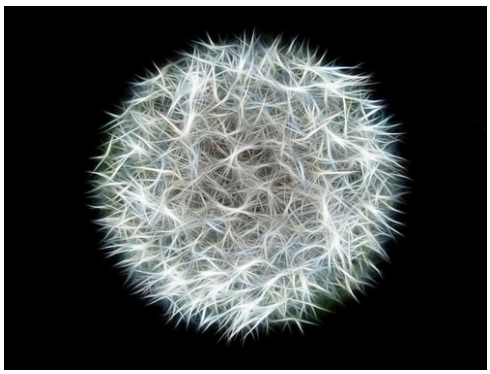
Franz’s Knowledge Graph Solution includes both technology and services for building industrial strength Entity-Event Knowledge Graphs based on best-of-class tools, products, knowledge, skills and experience. At the core of the solution is Franz’s graph database technology, AllegroGraph with FedShard, which is utilized by dozens of the top F500 companies worldwide and enables businesses to extract sophisticated decision insights and predictive analytics from highly complex, distributed data that cannot be uncovered with conventional databases.

Franz delivers the expertise for designing ontology and taxonomy-based solutions by utilizing standards-based development processes and tools. Franz also offers data integration services from siloed data using W3C industry standard semantics, which can then be continually integrated with information that comes from other data sources. In

addition, the Franz data science team provides expertise in custom algorithms to maximize data analytics and uncover hidden knowledge.

Ubiquitous AI Demands A New Type Of Database Sharding

Forbes published the following article by Dr. Jans Aasman, Franz Inc.’s CEO.



The notion of sharding has become increasingly crucial for selecting and optimizing database architectures. In many cases, sharding is a means of horizontally distributing data; if properly implemented, it results in near-infinite scalability. This option enables database availability for business continuity, allowing organizations to replicate databases among geographic locations. It’s equally useful for load balancing, in which computational necessities (like processing) shift between machines to improve IT resource allocation.

However, these use cases fail to actualize sharding’s full potential to maximize database performance in today’s post-big data landscape. There’s an even more powerful form of sharding, called “hybrid sharding,” that drastically improves the speed of query results and duly expands the complexity of the questions that can be asked and answered. Hybrid sharding is the ability to combine data that can be partitioned into

shards with data that represents knowledge that is usually un-shardable.

This hybrid sharding works particularly well with the knowledge graph phenomenon leveraged by the world's top data-driven companies. Hybrid sharding also creates the enterprise scalability to query scores of internal and external sources for nuanced, detailed results, with responsiveness commensurate to that of the contemporary AI age.

Read the full article at Forbes.

Natural Language Processing and Machine Learning in AllegroGraph

The majority of our customers build Knowledge Graphs with Natural Language and Machine learning components. Because of this trend AllegroGraph now offers strong support for the use of Natural Language Processing and Machine learning.

Franz Inc has a team of NLP engineers and Taxonomy experts that can help with building turn-key solutions. In general however, our customers already have some expertise in house. In those cases we train customers in how to take the output of NLP and ML processing and turn that into an efficient Knowledge Graph based on best practices in the industry.

This document primarily describes the NLP and ML plug-in AllegroGraph.

Note that many enterprises already have a data science team with NLP experts that use modern open source NLP tools like Spacy, Gensim or Polyglot, or Machine Learning based NLP tools like BERT and Scikit-Learn. In another blog about Document Handling we describe a pipeline of how to deal with NLP in Document Knowledge Graphs by using our NLP and ML plugin and mix that with open source tools.

PlugIn features for Natural Language Processing and Machine Learning in AllegroGraph.

Here is the outline of the plugin features that we are going to describe in more detail.

Machine learning

- data acquisition
- classifier training
- feature extraction support
- performance analysis
- model persistence

NLP

- handling languages
- handling dictionaries
- tokenization
- entity extraction
- Sentiment analysis
- basic pattern matching

SPARQL Access

- Future development

Machine Learning

ML: Data Acquisition

Given that the NLP and ML functions operate within AllegroGraph, after loading the plugins, data acquisition can be performed directly from the triple-store, which drastically simplifies the data scientist workflow. However, if the data is not in AllegroGraph yet we can also import it directly from ten formats of triples or we can use our additional capabilities to import from CSV/JSON/JSON-LD.

Part of the Data Acquisition is also that we need to pre-process the data for training so we provide these three functions:

- prepare-training-data
- split-dev-test
- equalize (for resampling)

Machine Learning: Classifiers

- Currently we provide simple linear classifiers. In case there's a need for neural net or other advanced classifiers, those can be integrated on-demand.
- We also provide support for online learning (online machine learning is an ML method in which data becomes available in a sequential order and is used to update the best predictor for future data at each step, as opposed to batch learning techniques which generate the best predictor by learning on the entire training data set at once). This feature is useful for many real-world data sets that are constantly updated.
- The default classifiers available are Averaged Perceptron and AROW

Machine Learning: Feature Extraction

Each classifier is expecting a vector of features: either feature indices (indicative features) or pairs of numbers (index – value). These are obtained in a two-step process:

1. A classifier-specific extract-features method should be defined that will return raw feature vector with features identified by strings of the following form: prefix|feature.

The prefix should be provided as a keyword argument to the collect-features method call, and it is used to distinguish similar features from different sources (for instance, for distinct predicates).

2. Those features will be automatically transformed to unique integer ids. The resulting feature vector of indicator features may look like the following: #(1 123 2999 …)

Note that these features may be persisted to AllegroGraph for repeated re-use (e.g. for experimenting with classifier hyperparameter tuning or different classification models).

Many possible features may be extracted from data, but there is a set of common ones, such as:

1. individual tokens of the text field
2. ngrams (of a specified order) of the text field
3. presence of a token in a specific dictionary (like, the dictionary of slang words)
4. presence/value of a certain predicate for the subject of the current triple
5. length of the text

And in case the user has a need for special types of tokens we can write specific token methods, here is an example (in Lisp) that produces an indicator feature of a presence of emojis in the text:

```
(defmethod collect-features ((method (eql :emoji)) toks &key pred)
(dolist (tok toks)
(when (some #'(lambda (code)
  (or (<= #x1F600 code #x1F64F)
```

```
      (<= #x1F650 code #x1F67F)
      (<= #x1F680 code #x1F6FF)))
   (map 'vector #'char-code tok))
(return (list "emoji")))))
```

Machine Learning: Integration with Spacy

The NLP and ML community invents new features and capabilities at an incredible speed. Way faster than any database company can keep up with. So why not embrace that? Whenever we need something that we don’t have in AllegroGraph yet we can call out to Spacy or any other external NLP tool. Here is an example of using feature extraction from Spacy to collect indicator features of the text dependency parse relations:

```
(defmethod collect-features ((method (eql :dep)) deps &key
pred dep-type dep-labels)
 (loop :for ds :in deps :nconc
  (loop :for dep :in ds
   :when (and (member (dep-tag dep) dep-labels)
              (dep-head dep)
              (dep-tok dep))
    :collect (format nil "dep|~a|~a_~a"
              dep-type
              (tok-word (dep-head dep)
              (tok-word (dep-tok dep))))))
```

The demonstrated integration uses Spacy Docker instance and its HTTP API.

Machine Learning: Classifier Analysis

We provide all the basic tools and metrics for classifier quality analysis:

- accuracy
- f1, precision, recall
- confusion matrix
- and an aggregated classification report

Machine Learning: Model Persistence

The idea behind model persistence is that all the data can be stored in AllegroGraph, including features and classifier models. AllegroGraph stores classifiers directly as triples. This is a far more robust and language-independent approach than currently popular among data scientists reliance on Python pickle files. For the storage we provide a basic triple-based format, so it is also possible to interchange the models using standard RDF data formats.

The biggest advantage of this approach is that when adding text to AllegroGraph we don't have to move the data externally to perform the classification but can keep the whole pipeline entirely internal.

Natural Language Procession (NLP)

NLP: Language Packs

Most of the NLP tools are language-dependent: i.e. there's a general function that uses language-specific model/rules/etc. In AllegroGraph, support for particular languages is provided on-demand and all the language-specific is grouped in the so called "language pack" or langpack, for short – a directory with a number of text and binary files with predefined names.

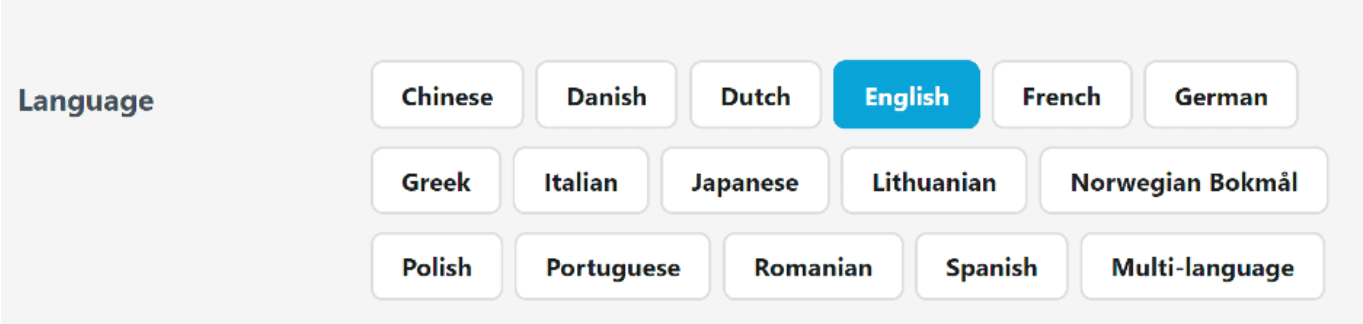
Currently, the langpack for English is provided at nlp/langs/en.zip, with the following files:

- contractions.txt – a dictionary of contractions
- abbrs.txt – a dictionary of abbreviations
- stopwords.txt – a dictionary of stopwords
- pos-dict.txt – positive sentiment words
- neg-dict.txt – negative sentiment words
- word-tok.txt – a list of word tokenization rules

Additionally, we use a general dictionary, a word-form dictionary (obtained from Wiktionary), and custom lexicons.

Loading a langpack for a particular language is performed using load-langpack.

Creating a langpack is just a matter of adding the properly named files to the directory and can be done manually. The names of the files should correspond to the names of the dictionary variables that will be filled by the pack. The dictionaries that don’t have a corresponding file will be just skipped.We have just finished creating a langpack for Spanish and it will be published soon. In case you need other dictionaries we use our AG/Spacy infrastructure. Spacy recently added a comprehensive list of new languages:



NLP: Dictionaries

Dictionaries are read from the language packs or other sources and are kept in memory as language-specific hash-tables. Alongside support for storing the dictionaries as text files, there are also utilities for working with them as triples and putting them into the triple store.

Note that we at Franz Inc specialize in Taxonomy Building using various commercial taxonomy building tools. All these tools can now export these taxonomies as a mix of SKOS taxonomies and OWL. We have several functions to read directly from these SKOS taxonomies and turn them into dictionaries

that support efficient phrase-level lookup.

NLP: Tokenization

Tokenization is performed using a time-proven rule-based approach. There are 3 levels of tokenization that have both a corresponding specific utility function and an :output format of the tokenize function:

:parags – splits the text into a list of lists of tokens for paragraphs and sentences in each paragraph
:sents – splits the text into a list of tokens for each sentence
:words – splits the text into a plain list of tokens

Paragraph-level tokenization considers newlines as paragraph delimiters. Sentence-level tokenization is geared towards western-style writing that uses dot and other punctuation marks to delimit sentences. It is, currently, hard-coded, but if the need arises, additional handling may be added for other writing systems. Word-level tokenization is performed using a language-specific set of rules.

NLP: Entity Extraction

Entity extraction is performed by efficient matching (exactly or fuzzy) of the token sequences to the existing dictionary structure.

It is expected that the entities come from the triple store and there’s a special utility function that builds lookup dictionaries from all the triples of the repository identified by certain graphs that have a skos:prefLabel or skos:altLabel property. The lookup may be case-insensitive with the exception of abbreviations (default) or case-sensitive.

Similar to entity extraction, there’s also support for spotting sentiment words. It is performed using the positive/negative words dictionaries from the langpack.

One feature that we needed to develop for our customers is ‘heuristic entity extraction’ . In case you want to extract complicated product names from text or call-center conversations between customers and agents you run into the problem that it becomes very expensive to develop altLabels in a taxonomy tool. We created special software to facilitate the automatic creation of altlabels.

NLP: Basic Pattern Matching for relationship and event detection

Getting entities out of text is now well understood and supported by the software community. However, to find complex concepts or relationships between entities or even events is way harder and requires a flexible rule-based pattern matcher. Given our long time background in Lisp and Prolog one can imagine we created a very powerful pattern matcher.

SPARQL Access

Currently all the features above can be controlled as stored procedures or using Lisp as the command language. We have a new (beta) version that uses SPARQL for most of the control. Here are some examples. Note that fai is a magic-property namespace for “AI”-related stuff and inc is a custom namespace of an imaginary client:

1. Entity extraction

```
select ?ent {
   ?subj fai:entityTaxonomy inc:products .
   ?subj fai:entityTaxonomy inc:salesTerms .
   ?subj fai:textPredicate inc:text .
     ?subj fai:entity(fai:language "en", fai:taxonomy
inc:products) ?ent .
}
```

The expressions ?subj fai:entityTaxonomy inc:poducts and ?subj fai:entityTaxonomy inc:salesTerms specify which taxonomies to use (the appropriate matchers are cached).

The expression ?subj fai:entity ?ent will either return the already extracted entities with the specified predicate (fai:entity) or extract the new entities according to the taxonomies in the texts accessible by fai:textPredicate.

2. fai:sentiment will return a single triple with sentiment score:

```
select ?sentiment {
   ?subj fai:textPredicate inc:text .
   ?subj fai:sentiment ?sentiment .
   ?subj fai:language "en" .
   ?subj fai:sentimentTaxonomy franz:sentiwords .
}
```

3. Text classification:
Provided inc:customClassifier was already trained previously, this query will return labels for all texts as a result of classification.

```
select ?label {
?subj fai:textPredicate inc:text .
?subj fai:classifier inc:customClassifier .
?subj fai:classify ?label .
?label fai:storeResultPredicate inc:label .
}
```

Further Development
Our team is currently working on these new features:

- A more accessible UI (python client & web) to facilitate NLP and ML pipelines
- Addition of various classifier models
- Sequence classification support (already implemented for a customer project)
- Pre-trained models shipped with AllegroGraph (e.g. English NER)

- Graph ML algorithms (deepwalk, Google Expander)
- Clustering algorithms (k-means, OPTICS)

Document Knowledge Graphs with NLP and ML

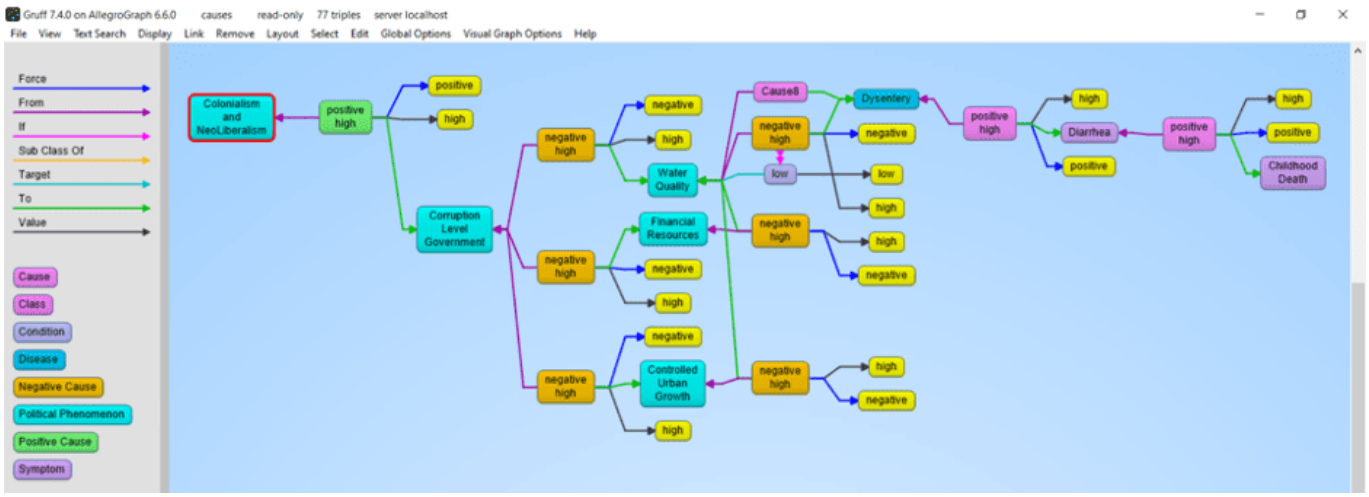
A core competency for Franz Inc is turning text and documents into Knowledge Graphs (KG) using Natural Language Processing (NLP) and Machine Learning (ML) techniques in combination with AllegroGraph. In this document we discuss how the techniques described in [NLP and ML components of AllegroGraph] can be combined with popular software tools to create a robust Document Knowledge Graph pipeline.

We have applied these techniques for several Knowledge Graphs but in this document we will primarily focus on three completely different examples that we summarize below. First is the Chomsky Legacy Project where we have a large set of very dense documents and very different knowledge sources, Second is a knowledge graph for an intelligent call center where we have to deal with high volume dynamic data and real-time decision support and finally, a large government organization where it is very important that people can do a semantic search against documents and policies that steadily change over time and where it is important that you can see the history of documents and policies.

Example [1] Chomsky Knowledge Graph

The Chomsky Legacy Project is a project run by a group of admirers of Noam Chomsky with the primary goal to preserve all his written work, including all his books, papers and interviews but also everything written about him. Ultimately students, researchers, journalists, lobbyists, people from the AI community, and linguists can all use this knowledge graph for their particular goals and questions.

The biggest challenges for this project are finding causal relationships in his work using event and relationship extraction. A simple example we extracted from an author quoting Chomsky is that neoliberalism ultimately causes childhood death.



Example 2: N3 Results and the Intelligent Call Center

This is a completely different use case (See a recent KMWorld Articlehttps://allegrograph.com/knowledge-graphs-enhance-customer-experience-through-speed-and-accuracy/). Whereas the previous use case was very static, this one is highly dynamic. We analyze in real-time the text chats and spoken conversations between call center agents and customers. Our knowledge graph software provides real-time decision support to make the call center agents more efficient. N3 Results helps big tech companies to sell their high tech solutions, mostly cloud-based products and services but also helps their clients sell many other technologies and services.

The main challenge we tackle is to really deeply understand what the customer and agent are talking about. None of this can be solved by only simple entity extraction but requires elaborate rule-based and machine learning techniques. Just to give a few examples. We want to know if the agent talked about their most important talking points: that is, did the agent ask if the customer has a budget, or the authority to make a decision or a timeline about when they need the new technology or whether they actually have expressed their need. But also whether the agent reached the right person, and whether the agent talked about the follow-up. In addition, if the customer talks about competing technology we need to recognize that and provide the agent in real-time with a battle card specific to the competing technology. And in order to be able to do the latter, we also analyzed the complicated marketing materials of the clients of N3.

Example 3: Complex Government Documents
Imagine a regulatory body with tens of thousands of documents. Where nearly every paragraph has reference to other paragraphs in the same document or other documents and the documents change over time. The goal here is to provide the end-users in the government with the right document given their current task at hand. The second goal is to keep track of all the changes in the documents (and the relationship between documents) over time.

The Document to Knowledge Graph Pipeline

Process Name	Input	Output
1. Custom Taxonomy Creation	Corpus Analytics, Taxonomy tool	A SKOS taxonomy containing concepts, concept hierarchy, prefLabels, altLabels.
2. Document Preparation	Documents (pdf, word, ppt, xlsx), Apache Tika, Spacy for XML cleanup	An XML version of each document
3. Extract Document Meta Data	Document + Apache Tika	JSON dictionary of the Document MetaData
4. XML-to-Triples	XML+JSON dictionary, XMLToTriples.py	Graph-based document tree with chapters, sections, and paragraphs as triples. Also includes meta data as triples
5. Entity-Extraction	Paragraphs + taxonomies + AllegroGraph Entity extract or external extractors	Concepts, persons, places, currencies. Connected to paragraphs
6. LOD Enrichment	Paragraphs + IBM Natural Language Understanding.	Concept categories and links to DBpedia and GeoNames, etc.
7. Complex Relationship and Event extraction.	Paragraphs + Taxonomy + Rules in Spacy or AllegroGraph	Complex events and relationships, References to other document sections.
8. NLP and ML	Chapters and paragraphs + all the tools described [here], but also using Spacy, Gensim, BERT, SciKit Learn.	Similarities, sentiment, query answering, smart search, text classification, word embeddings, abstracts
9. Versioning and Document tracking	Old + New document, compare.py	Old document in historic repository, new document in current, changed graph.
10. Statistical Relationships	Concepts + OddRatio.py or OddsRatio.cl	Statistical relationships between concepts.

Let us first give a quick summary in words of how we turn documents into a Knowledge Graph.

[1] Taxonomy Creation

Taxonomy of all the concepts important to the business using open source or commercial taxonomy builders. An available industry taxonomy is a good starting point for additional customizations.

[2] Document Preparation

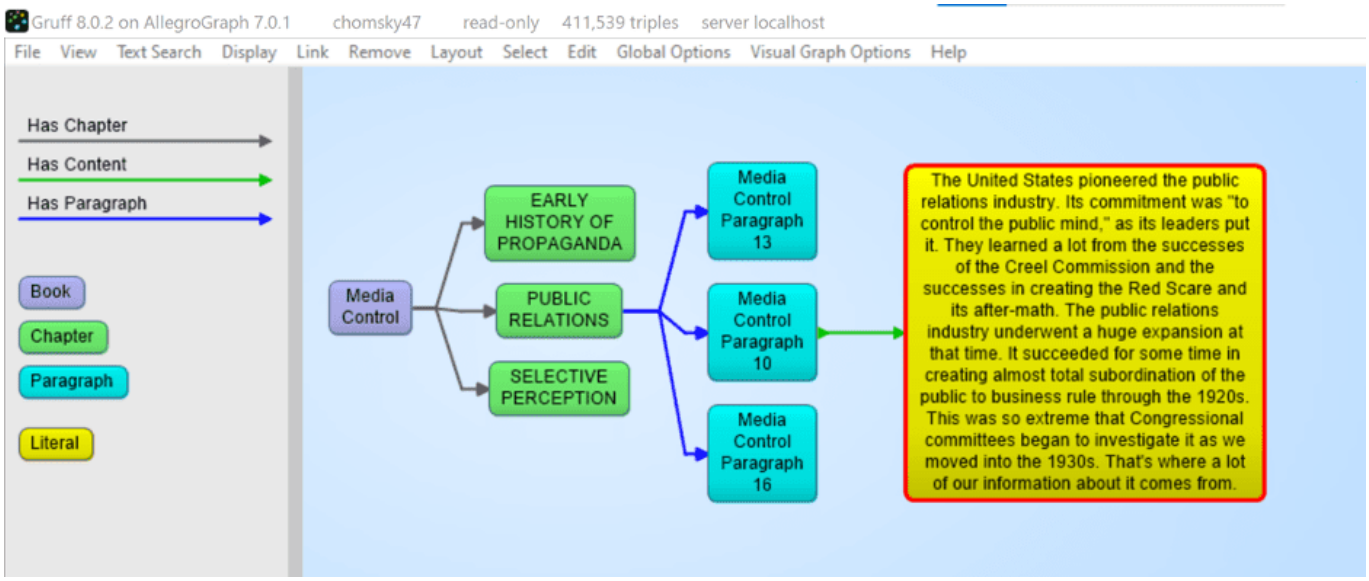
We then take a document and turn it into an intermediate XML using Apache Tika. Apache Tika supports more than 1000 document types and although Apache Tika is a fantastic tool, the output is still usually not clean enough to create a graph from, so we use Spacy rules to clean up the XML to make it as uniform as possible.

[3] Extract Document MetaData

Most documents also contain document metadata (author, date, version, title, etc) and Apache Tika will also deliver the metadata for a document as a JSON object.

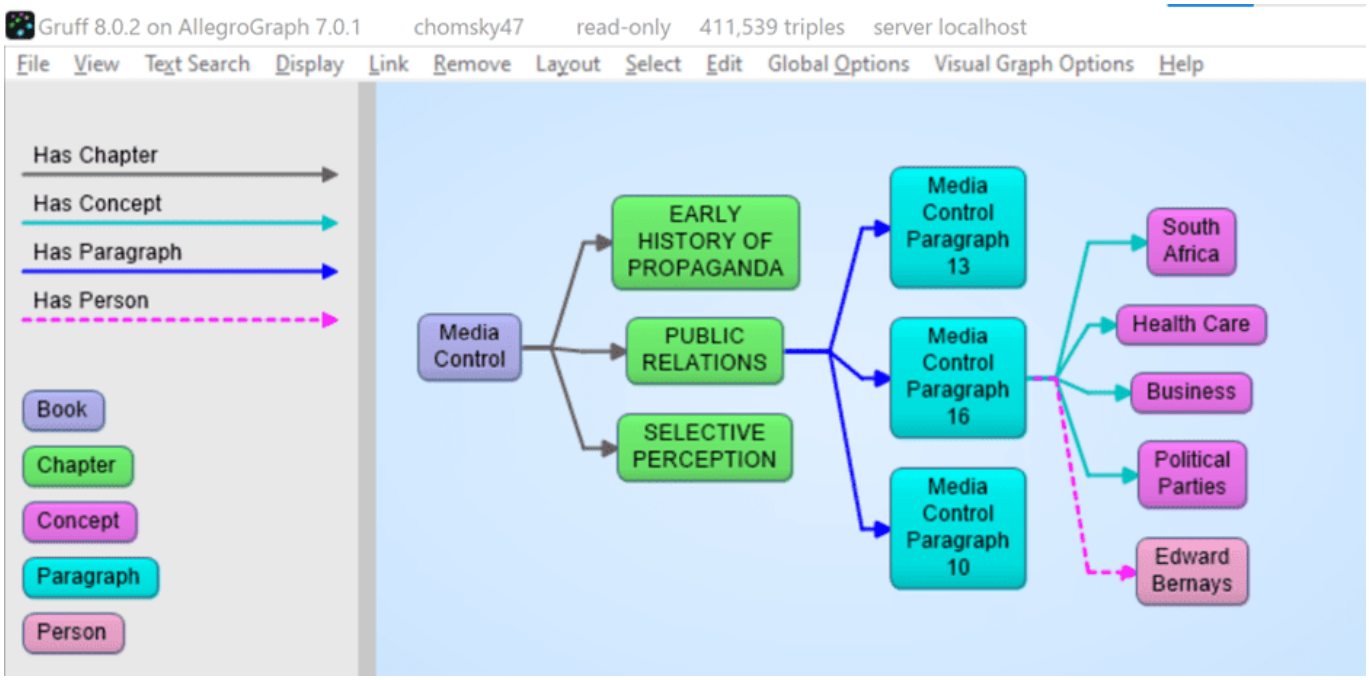
[4] XML to Triples

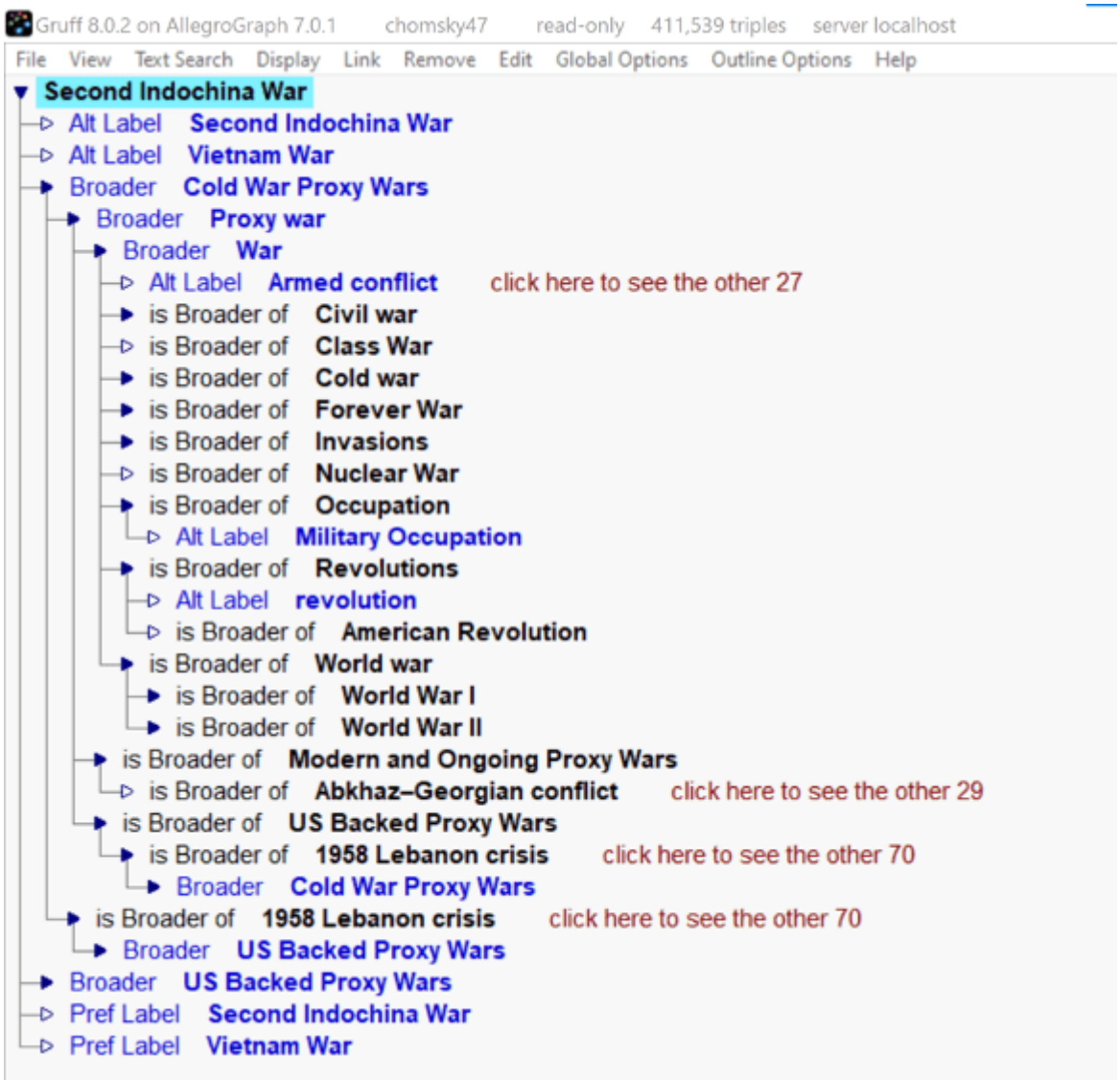
Our tools ingest the XML and metadata and transform that into a graph-based document tree. The document is the root and from that, it branches out into chapters, optionally sections, all the way down to paragraphs. The ultimate text content is in the paragraphs. In the following example we took the XML version of Noam Chomsky's book Media Control and turned that into a tree. The following shows a tiny part of that tree. We start with the Media Control node, then we show three (of the 11) chapters, for one chapter we show three (of the 6) paragraphs, and then we show the actual text in that paragraph. We sometimes can go even deeper to the level of sentences and tokens but for most projects that is overkill.



[5] Entity Extractor

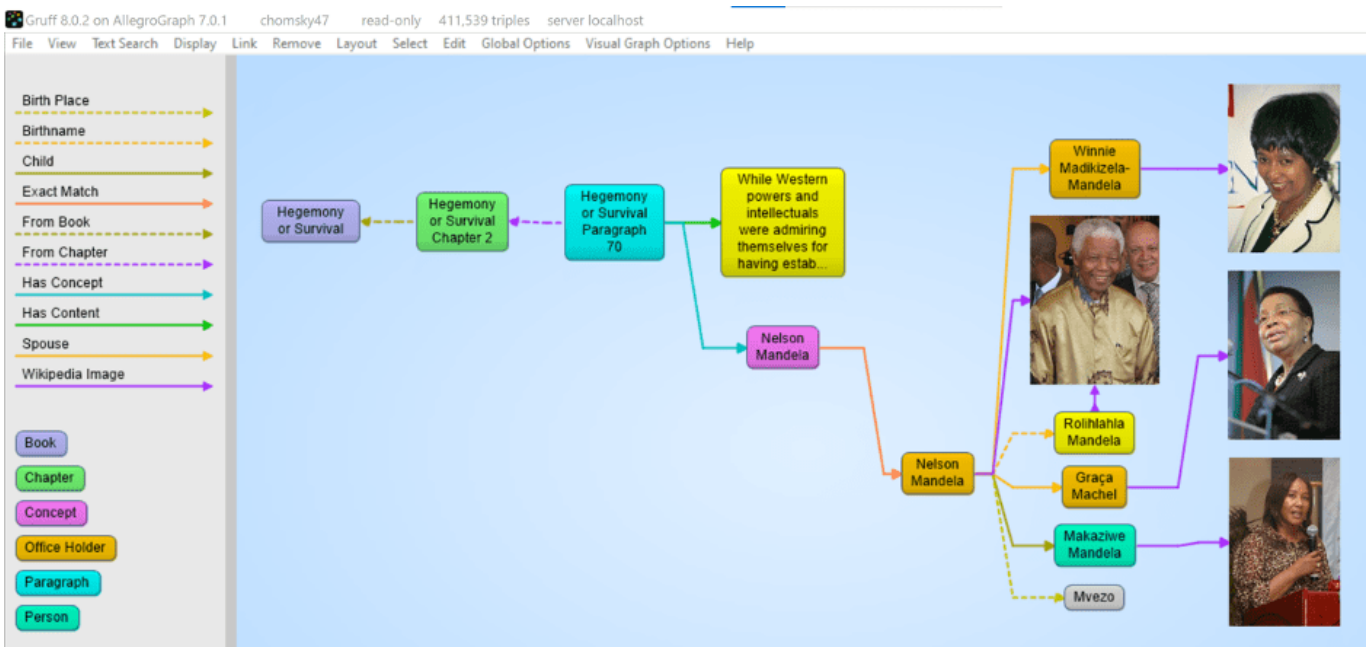
AllegroGraph's entity extractor takes as input the text of each paragraph in the document tree and one or more of the taxonomies and returns recognized SKOS concepts based on prefLabels and altLabels. AllegroGraph's entity extractor is state of the art and especially powerful when it comes to complex terms like product names. We find that in our call center a technical product name can sometimes have up to six synonyms or very specific jargon. For example the Cisco product Catalyst 9000 will also be abbreviated as the cat 9k. Instead of developing altLabels for every possible permutation that human beings *will* use, we have specialized heuristics to optimize the yield from the entity extractor. The following picture shows 4 (of the 14) concepts discovered in paragraph 16. Plus one person that was extracted by IBM's NLU.





[6] Linked Data Enrichment

In many use cases, AllegroGraph can link extracted entities to concepts in the linked data cloud. The most prominent being DBpedia, wikidata, the census database, GeoNames, but also many Linked Open Data repositories. One tool that is very useful for this is IBM's Natural Language Understanding program but there are others available. In the following image we see that the Nelson Mandela entity (Red) is linked to the dbpedia entity for Nelson Mandela and that then links to the DBpedia itself. We extracted some of his spouses and a child with their pictures.



[7] Complex Relationship and Event Extraction

Entity extraction is a first good step to 'see' what is in your documents but it is just the first step. For example: how do you find in a text whether company C1 merged with company C2. There are many different ways to express the fact that a company fired a CEO. For example: Uber got rid of Kalanick, Uber and Kalanick parted ways, the board of Uber kicked out the CEO, etc. We need to write explicit symbolic rules for this or we need a lot of training data to feed a machine learning algorithm.

[8] NLP and Machine Learning

There are many many AI algorithms that can be applied in Document Knowledge Graphs. We provide best practices for topics like:

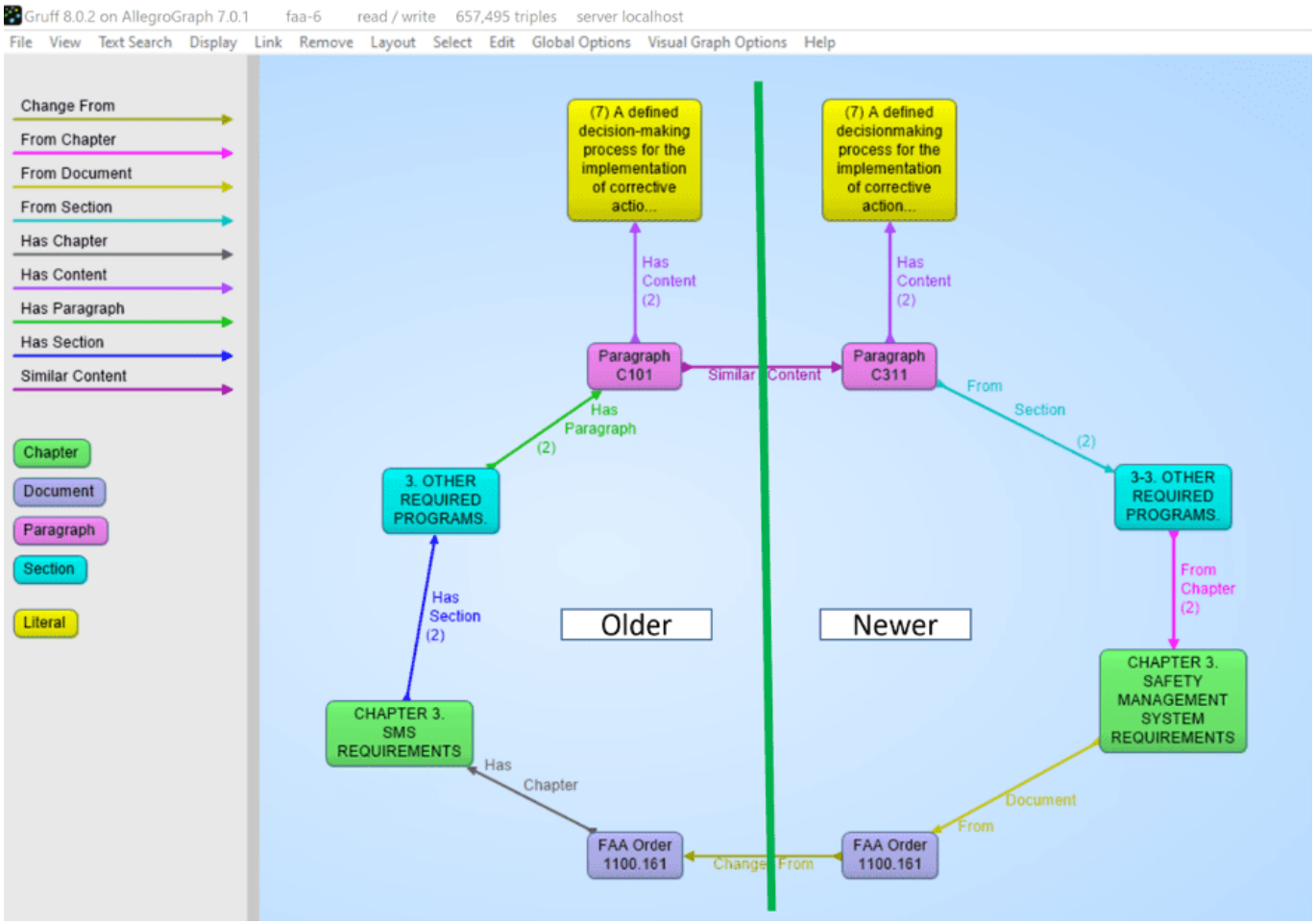
[a] Sentiment Analysis, using good/bad word lists or training data.
[b] Paragraph or Chapter similarity using statistical techniques like Gensim similarity or symbolic techniques where we just the overlap of recognized entities as a function of the size of a text.
[c] Query answering using word2vec or more advanced techniques like BERT
[d] Semantic search using the hierarchy in SKOS taxonomies.
[e] Summarization techniques for Abstractive or Extractive abstracts using Gensim or Spacy.

[9] Versioning and Document tracking

Several of our customers with Document Knowledge Graphs have noted the one constant in all of these KGs is that documents change over time. As part of our solution, we have created best practices where we deal with these changes. A crucial first step is to put each document in its own graph (i.e. the fourth element of every triple in the document tree is the document id itself). When we get a new version of a document the document ID changes but the new document will point back to the old version. We then compute which paragraphs stayed the same within a certain margin (there are always changes in whitespace) and we materialize what paragraphs disappeared in the new version and what new paragraphs appeared compared to the previous version. Part of the best practice is to put the old version of a document in a historical database that at all times can be federated with the ‘current’ set of documents.

Note that in the following picture we see the progression of a document. On the right hand side we have a newer version of a document 1100.161 with a chapter -> section -> paragraph -> contents where the content is almost the same as the one in

the older version. But note that the newer one spells ‘decision making’ as one word whereas the older version said ‘decision-making’. Note that also the chapter titles and the section titles are almost the same but not entirely. Also, note that the new version has a back-pointer (changed-from) to the older version.



[10] Statistical Relationships

One important analytic one can do on documents is to look at the co-occurrence of terms. Although, given that certain words might occur more frequently in text, we have to correct the co-occurrence between words for the frequency of the two terms in a co-occurrence to get a better idea of the ‘surprisingness’ of a co-occurrence. The platform offers several techniques in Python and Lisp to compute these co-occurrences. Note that in the following picture we computed the odds ratios between recognized entities and so we see in

the following gruff picture that if Noam Chomsky talks about South Africa then the chances are very high he will also talk about Nelson Mandela.